NeuroSymb-MRG: Differentiable Abductive Reasoning with Active Uncertainty Minimization for Radiology Report Generation

Rong Fu[1⋆], Yiqing Lyu[2†], Chunlei Meng[3], Muge Qi[4], Yabin Jin[5,6,7⋆], Qi Zhao[1], Li Bao[8], Juntao Gao[2], Fuqian Shi[9], Nilanjan Dey[10], Wei Luo[7], Wenxin Zhang[11], and Simon Fong[1]

[1] University of Macau, Macau, China
[2] Tsinghua University, Beijing, China
[3] Fudan University, Shanghai, China
[4] Peking University, Beijing, China
[5] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
[6] Shunde Innovation School, University of Science and Technology Beijing, Foshan, China
[7] The First People's Hospital of Foshan, Foshan, China
[8] Capital Medical University, Beijing, China
[9] NYU Langone Health, NY, USA
[10] Techno International New Town, Kolkata, India
[11] University of Chinese Academy of Sciences, Beijing, China
{mc46603@um.edu.mo, lvyq22@mails.tsinghua.edu.cn, jinyabin1990@qq.com}

Abstract. Automatic generation of radiology reports seeks to reduce clinician workload while improving documentation consistency. Existing methods that adopt encoder-decoder or retrieval-augmented pipelines achieve progress in fluency but remain vulnerable to visual-linguistic biases, **structural misalignment**, and lack of explicit multi-hop clinical reasoning. We present NeuroSymb-MRG, a unified framework that integrates NeuroSymbolic abductive reasoning with active uncertainty minimization to produce structured, **interpretable** reports. The system maps image features to probabilistic clinical concepts, composes differentiable logic-based reasoning chains, decodes those chains into templated clauses, and refines the textual output via retrieval and constrained language-model editing. An active sampling loop driven by rule-level uncertainty and diversity guides targeted training optimization and promptbook refinement. Experiments on standard benchmarks demonstrate consistent improvements in structural coherence and interpretability and standard language metrics compared to representative baselines.

Keywords: NeuroSymbolic report generation, differentiable abductive reasoning, uncertainty minimization, retrieval augmented generation

⋆ Rong Fu and Yiqing Lyu† contribute equally to this work. Corresponding authors: Rong Fu (mc46603@um.edu.mo) and Yabin Jin (jinyabin1990@qq.com).

1 Introduction

Radiology reports are essential for documenting imaging findings and clinical impressions, yet manual authoring is time-consuming and scales poorly, motivating automated report generation [26,25]. Early work adapted encoder–decoder captioning with attention, followed by spatially aware mechanisms, memory and meshed-memory modules, and retrieval- or prototype-based fusion [20,2,6,5,4,27], while recent systems leverage frozen large language models with lightweight visual alignment and prompt tuning [28,14]. Despite gains in fluency, public datasets such as MIMIC-CXR and IU X-ray introduce distributional imbalance and label noise that hinder reliable end-to-end learning [12,8]. Persistent challenges include hallucinated but fluent statements from data-limited text decoders, the absence of explicit multi-hop reasoning for clause-level justifications, and poor coverage of low-prevalence conditions due to imbalance and rarity. Prior remedies using cross-modal causal interventions, memory networks, and diagnosis-driven prompts improve evidence use but do not yield integrated, human-interpretable reasoning chains with rule-level uncertainty quantification [18,3,11,9]. We address these gaps by combining differentiable neuro-symbolic reasoning with retrieval-augmented generation and an uncertainty-aware active sampling loop, enabling end-to-end optimization with decodable hard rules for explanation and directing clinician effort toward ambiguous, high-impact cases for improved data efficiency and factual reliability [15,1,30,13,16].

Our contributions are as follows. First, we formulate radiology report generation as a structured mapping from image to a triple of probabilistic concepts, a differentiable abductive reasoning chain, and a templated textual draft, and we implement a differentiable logic layer that composes soft logical operators into multi-hop rules suitable for medical claims. Second, we design a hybrid generation pipeline that decodes soft rule activations into canonical clause templates, augments these drafts with retrieval evidence and constrained LLM paraphrasing, and employs a verifier to detect high-risk contradictions. Third, we introduce an active uncertainty minimization mechanism that measures rule-level predictive entropy with Monte Carlo dropout, selects diverse high-entropy cases using a k-center strategy, and uses a learned feedback simulator to accelerate targeted training optimization. Finally, we validate the design on standard benchmarks and show that integrating explicit reasoning and rule-level active sampling improves structural coherence and interpretability and common language metrics relative to strong baselines. The experimental setup, ablations, and implementation details are informed by common practices and datasets in the field.

2 Related Work

2.1 Radiology Report Generation Strategies

Radiology report generation has progressed from adapting generic encoder–decoder captioning with attention [26,25] to domain-specific spatial mechanisms such as

adaptive sentinel control and bottom-up attention [20,2], followed by Transformer-based meshed-memory and relational architectures with cross-modal memory and hierarchical alignment that better connect visual findings to anatomical semantics [6,5,4,31]. Subsequent advances improved diagnostic reliability via progressive generation and reinforced alignment [21,22] and increasingly employed large language models through prompt tuning and frozen-weight designs [14,28].

2.2 Neural-Symbolic AI and Differentiable Reasoning

Neural–symbolic methods combine statistical learning with explicit logical structure to improve interpretability in clinical reasoning [15], leveraging differentiable formulations such as fuzzy Zadeh T-norms to build learnable logic-gate networks for gradient-based Boolean composition [30] and enabling end-to-end reasoning in structured tasks like trajectory prediction [1]. Higher-level frameworks addressing abductive inference [7] further demonstrate their capacity for interpretable medical explanation, with logical constraints preserved throughout training to maintain faithfulness beyond black-box models.

2.3 Uncertainty Estimation and Active Learning

Bayesian dropout provides reliable predictive uncertainty in high-risk tasks such as seismic inversion [13] and medical diagnostics, enabling hallucination mitigation and targeted supervision. These signals also support active learning strategies that pair uncertainty with diversity-based selection [16], while broader frameworks in multi-agent reinforcement learning and retrieval-augmented structuring [10] motivate our use of active uncertainty reduction within the neuro-symbolic reasoning module.

2.4 Pre-training and Knowledge Integration

Modern medical vision systems build on large-scale datasets such as MIMIC-CXR [12] and curated radiology collections [8], with self-supervised learning [23] and knowledge integration techniques [18,3] enabling robust, evidence-grounded representations. Additional advances including medical semantic augmentation [29], prototype-based modeling [27], graph-guided semi-supervision [32] and diagnosis-aware prompting [11] further strengthen consistency and align data-driven predictions with expert reasoning.

3 Methodology

3.1 Formal Problem Characterization

We define the diagnostic task as a mapping $\mathcal{M}$ from a high-dimensional radiological image $I \in \mathbb{R}^{H\times W\times C}$ to a structured interpretative tuple $\widehat{R} = (\widehat{C}, \widehat{\mathcal{L}}, \widehat{T})$. In this framework, $\widehat{C}$ represents a set of probabilistic clinical concept activations, $\widehat{\mathcal{L}}$

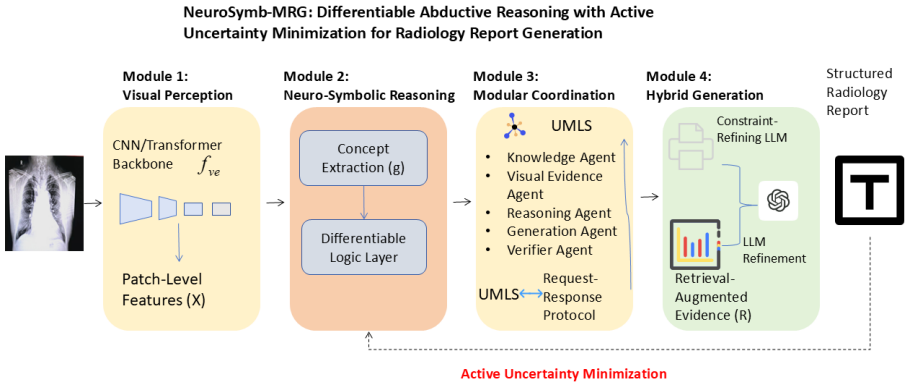

Fig. 1: Architectural overview of the **NeuroSymb-MRG** framework for transparent and clinically grounded radiology report generation.

signifies a multi-step neuro-symbolic reasoning trajectory, and $\widehat{T}$ denotes the synthesized textual report. The objective is to optimize $\mathcal{M}$ such that the generated $\widehat{R}$ aligns with the ground-truth report $R^{\star}$ through the following transformation:

$$\mathcal{M} : I \mapsto \widehat{R} = \left(\widehat{C}, \widehat{\mathcal{L}}, \widehat{T}\right) \tag{1}$$

where I serves as the visual input and $\widehat{R}$ encapsulates the clinical findings and logic.

3.2 Architectural Framework

The proposed architecture synchronizes five modular agents to convert raw pixels into clinical narratives. This pipeline integrates self-supervised visual encoding, differentiable logic for transparency, and retrieval-augmented generation under an automated orchestration protocol.

3.3 Visual Representation and Feature Projection

A self-supervised visual backbone f_{ve} extracts localized patch features $X \in \mathbb{R}^{M \times C'}$ from the input image, formulated as:

$$X = f_{\text{ve}}(I) \tag{2}$$

where M denotes the spatial patch count and C' indicates the feature dimensionality. To align these representations with the language model's latent space, a linear projection f_{proj} computes visual tokens $V \in \mathbb{R}^{M \times D}$:

$$V = f_{\text{proj}}(X) \tag{3}$$

where D represents the embedding dimension of the language decoder.

3.4 Concept Prediction and Spatial Integration

The concept extraction module employs a transformer-based mechanism to capture inter-patch dependencies. The initial features X undergo self-attention within a transformer encoder block to yield refined features X':

$$X' = \text{TransEnc}(X; N_h = 8) \tag{4}$$

where N_h specifies the number of attention heads. Global context is then aggregated via average pooling to produce a vector $\bar{x} \in \mathbb{R}^{C'}$:

$$\bar{x} = \frac{1}{M} \sum_{m=1}^{M} X'_m \tag{5}$$

where X'_m refers to the m-th attended patch. This representation is fed into a multi-layer perceptron to derive K concept probabilities $\hat{c} \in [0,1]^K$:

$$\hat{c} = \sigma\big(W_2 \, \text{ReLU}(W_1 \bar{x} + b_1) + b_2\big) \tag{6}$$

where W and b are learnable weight matrices and bias vectors, and σ denotes the sigmoid activation.

3.5 Differentiable Neuro-Symbolic Logic

Transparent reasoning is achieved through a soft operator network r_θ that simulates Boolean algebra using t-norms. We define conjunction as the product $a \cdot b$, disjunction as the probabilistic sum $a + b - ab$, and negation as $1 - a$. Each diagnostic rule j is structured as a soft tree with learnable gating weights α:

$$\text{Node}(u, v; \alpha) = \alpha \cdot \text{AND}(u, v) + (1 - \alpha) \cdot \text{OR}(u, v) \tag{7}$$

where $\alpha \in [0,1]$ modulates the transition between logical operations. The activation for a specific rule j is expressed as:

$$\bar{r}_j = \Phi_j\big(\hat{c}; \theta_j\big) \tag{8}$$

where Φ_j represents the composed logic and θ_j denotes the internal parameters governing the soft tree structure.

3.6 Algorithmic Report Synthesis

The synthesis process converts rule activations into human-readable text by selecting templates from an inventory $\mathcal{T}$. A rule decoder d yields a sequence of clauses $\widehat{\mathcal{L}}$ through a slot-filling mechanism:

$$\widehat{\mathcal{L}} = d(\bar{r}) = \big(l_1, \ldots, l_H\big) \tag{9}$$

where l_i is a generated clause and H is the total count of prioritized rules. To enhance clinical accuracy, we retrieve N_r relevant fragments $\mathcal{R}$ from a vector database using the pooled visual embedding v:

$$\mathcal{R} = \text{Retrieve}(v; N_r) \tag{10}$$

where v is obtained by pooling visual tokens V. The final report $\widehat{T}_{\text{draft}}$ is then assembled by a filler function $\mathcal{F}$ that fuses retrieved evidence with decoded rules:

$$\widehat{T}_{\text{draft}} = \mathcal{F}\big(\widehat{\mathcal{L}}, \mathcal{R}, \mathcal{T}\big) \tag{11}$$

where $\mathcal{T}$ represents the pre-defined template skeletons.

3.7 Orchestration and Conflict Mitigation

The five modules operate in a fixed sequential pipeline: visual encoding, UMLS validation, logic composition, template filling, and contradiction detection. When conflicting claims arise for a specific slot s, a coordinator resolves the ambiguity by confidence-weighted voting over agent confidences γ_i:

$$x_s^\star = \arg\max_x \sum_{i:x_i=x} \gamma_i \tag{12}$$

where $x_s^\star$ is the consensus value and γ_i is the confidence score of the i-th agent. A knowledge agent adjusts γ_i via UMLS consistency checks; the pipeline is deterministic consensus, not learned negotiation.

3.8 Stochastic Uncertainty and Adaptive Sampling

To optimize the learning process, we quantify model uncertainty using predictive entropy $H[\bar{r}]$ derived from T_{MC} stochastic passes:

$$H[\bar{r}] = -\sum_{j=1}^{R} \bar{r}_j \log \bar{r}_j \tag{13}$$

where $\bar{r}$ is the mean activation across dropout samples and R is the rule cardinality. This metric guides an active sampling strategy that selects k samples by maximizing both uncertainty and feature diversity. A learned feedback simulator, implemented as a rule-level binary classifier trained on historical correction patterns, predicts whether a rule activation would be flagged incorrect, replacing human review during training.

3.9 Multi-Objective Optimization Strategy

The parameters are updated using a weighted composite loss $\mathcal{L}$ that balances reconstruction, concept accuracy, and rule fidelity:

$$\mathcal{L} = \sum_i \lambda_i \mathcal{L}_i + \lambda_\theta \|\theta\|^2 \tag{14}$$

where λ_i are task-specific coefficients. We implement homoscedastic uncertainty weighting to adapt these coefficients dynamically during training:

$$\tilde{\mathcal{L}} = \sum_i \frac{1}{2\sigma_i^2}\mathcal{L}_i + \log \sigma_i + \lambda_\theta \|\theta\|^2 \tag{15}$$

where σ_i is a learnable scalar representing the task-dependent noise level, ensuring no single objective dominates the gradient descent process.

Table 1: Quantitative comparison on IU X-ray and MIMIC-CXR, with best results in **bold** and second-best values underlined.

Method	IU X-ray[8]						MIMIC-CXR[12]					
	B-1	B-2	B-3	B-4	R-L	MTR	B-1	B-2	B-3	B-4	R-L	MTR
Show-Tell[26]	0.243	0.130	0.108	0.078	0.307	0.157	0.308	0.190	0.125	0.088	0.256	0.122
Transformer[25]	0.372	0.251	0.147	0.136	0.317	0.168	0.316	0.199	0.140	0.092	0.267	0.129
Att2in[24]	0.248	0.134	0.116	0.091	0.309	0.162	0.314	0.198	0.133	0.095	0.264	0.122
AdaAtt[20]	0.284	0.207	0.150	0.126	0.311	0.165	0.314	0.198	0.132	0.094	0.267	0.128
Up-Down[2]	–	–	–	–	–	–	0.317	0.195	0.130	0.092	0.267	0.128
M2Transformer[6]	0.402	0.284	0.168	0.143	0.328	0.170	0.332	0.210	0.142	0.101	0.264	0.134
R2Gen[5]	0.470	0.304	0.219	0.165	0.371	0.187	0.353	0.218	0.145	0.103	0.277	0.142
Contra.Attn.[19]	0.492	0.314	0.222	0.169	0.381	0.193	0.350	0.219	0.152	0.109	0.283	0.151
CMCL[17]	0.473	0.305	0.217	0.162	0.378	0.186	0.344	0.217	0.140	0.097	0.281	0.133
CMN[4]	0.475	0.309	0.222	0.170	0.375	0.191	0.353	0.218	0.148	0.106	0.278	0.142
Aligntransformer[31]	0.484	0.313	0.225	0.173	0.379	0.204	0.378	0.235	0.156	0.112	0.283	0.158
M2Tr.Prog.[21]	0.486	0.317	0.232	0.173	0.390	0.192	0.378	0.232	0.154	0.107	0.272	0.145
CMM+RL[22]	0.494	0.321	0.235	0.181	0.384	0.201	0.381	0.232	0.155	0.109	0.287	0.151
XPRONET*[27]	0.491	0.325	0.228	0.169	0.387	0.202	0.344	0.215	0.146	0.105	0.279	0.138
MCGN[29]	0.481	0.316	0.226	0.171	0.372	0.190	0.373	0.235	0.162	0.120	0.282	0.143
PPKED[18]	0.483	0.315	0.224	0.168	0.376	–	0.360	0.224	0.149	0.106	0.284	0.149
RAMT[32]	0.482	0.310	0.221	0.165	0.377	0.195	0.362	0.229	0.157	0.113	0.284	0.153
R2GenGPT[28]	0.482	0.306	0.215	0.158	0.370	0.200	0.387	0.248	0.170	0.123	0.280	0.149
VLCI†[3]	0.324	0.211	0.151	0.115	0.379	0.166	0.357	0.216	0.144	0.103	0.256	0.136
PromptMRG[11]	0.401	–	–	0.098	0.281	0.160	0.398	–	–	0.112	0.268	0.157
MedRAT[9]	0.455	–	–	0.129	0.349	–	0.365	–	–	0.086	0.251	–
MRG-LLM[14]	0.529	0.359	0.266	0.202	0.408	0.221	0.416	0.267	0.182	0.129	0.296	0.163
NeuroSymb-MRG (Ours)	**0.602**	**0.425**	**0.321**	**0.253**	**0.463**	**0.275**	**0.487**	**0.332**	**0.234**	**0.175**	**0.362**	**0.225**

4 Experiments

4.1 Datasets and evaluation metrics

We evaluate NeuroSymb-MRG on the MIMIC-CXR dataset [12], which provides 377,110 images and 227,835 reports, using the standard split of 368,960/2,991/5,159 images and 222,758/1,808/3,269 reports for training, validation and testing, and on IU X-ray [8] with 7,470 images and 3,955 reports under a patient-disjoint 7:1:2 split. Performance is quantified using BLEU-1 to BLEU-4, ROUGE-L, and METEOR.

Table 2: Ablation results for NeuroSymb-MRG on MIMIC-CXR, where ↓ denotes performance decreases relative to the full model.

Model Variant	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	Δ B-1
Full NeuroSymb-MRG	**0.487**	**0.332**	**0.234**	**0.175**	**0.362**	**0.225**	–
I. Core Architectural Components							
w/o Self-Supervised Visual Encoder (Random Init)	0.412	0.262	0.175	0.124	0.298	0.168	-15.4%
w/o Concept Attention Head (Raw Patch Features)	0.401	0.251	0.168	0.118	0.288	0.159	-17.7%
w/o Differentiable Logic Layer (End-to-End MLP)	0.428	0.278	0.190	0.136	0.315	0.185	-12.1%
w/o Rule-Guided Decoder (Direct LLM Generation)	0.439	0.289	0.198	0.144	0.321	0.191	-9.9%
II. Key Mechanisms (Novel Contributions)							
w/o Active Uncertainty Minimization (Random Sampling)	0.463	0.310	0.217	0.161	0.342	0.208	-4.9%
w/o Feedback Simulator (Direct Clinician Review)	0.476	0.324	0.228	0.170	0.354	0.218	-2.3%
w/o Modular Coordination (Single-Pipeline)	0.468	0.315	0.221	0.164	0.346	0.213	-3.9%
w/o Knowledge Agent (No UMLS Validation)	0.472	0.319	0.224	0.167	0.349	0.215	-3.1%
w/o Verifier Agent (No Contradiction Check)	0.475	0.323	0.228	0.170	0.353	0.217	-2.5%
w/o Dynamic Loss Weighting (Fixed λ)	0.471	0.318	0.222	0.165	0.348	0.213	-3.3%
III. Retrieval and Generation Components							
w/o Retrieval-Augmented Filling ($N_r = 0$)	0.459	0.305	0.216	0.161	0.339	0.206	-5.7%
w/o LLM Constrained Refinement (Template Only)	0.481	0.326	0.230	0.172	0.356	0.220	-1.2%
w/o Instance-Level Prompt Customization	0.445	0.292	0.203	0.150	0.328	0.198	-8.6%
IV. Baseline Comparisons							
Pure Neural Baseline (CNN+RNN)	0.398	0.248	0.165	0.116	0.285	0.156	-18.3%
Standard Transformer Baseline	0.416	0.268	0.182	0.129	0.302	0.171	-14.6%

4.2 Implementation details

The system uses a ConvNeXt-Tiny backbone pretrained on ImageNet-1K to extract 7×7 features ($M = 49$, $C' = 768$) projected into a $D = 4096$ token space, employs 50 learnable prompts, trains all modules with Adam (batch size 6, learning rate 1×10^{-4}) for 15 epochs on IU X-ray and 10 on MIMIC-CXR using PyTorch on four RTX 4090 GPUs, and performs active sampling with five MC-dropout passes and a per-round budget of 16 examples.

Table 3: Combined ablation results on the differentiable logic module, decoding strategies, training strategies, and retrieval settings on MIMIC-CXR.

Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
Full: Soft Tree (Prod. T-Norm + Prob. Sum + Gate)	**0.487**	**0.332**	**0.234**	**0.175**	**0.362**	**0.225**
Operator ablation: Min/Max T-norm / Conorm	0.479	0.325	0.228	0.170	0.355	0.219
Operator ablation: Łukasiewicz T-norm / Conorm	0.473	0.319	0.223	0.166	0.350	0.214
Operator ablation: Gödel T-norm / Conorm	0.468	0.315	0.219	0.162	0.345	0.210
Composition ablation: Fixed $\alpha = 1.0$ (AND only)	0.452	0.299	0.206	0.151	0.332	0.200
Composition ablation: Fixed $\alpha = 0.0$ (OR only)	0.461	0.308	0.214	0.158	0.340	0.207
MLP baseline (no symbolic structure)	0.441	0.290	0.200	0.147	0.325	0.195
Decoding ablation: no uncertainty qualifier	0.478	0.324	0.227	0.169	0.354	0.218
Decoding ablation: no weighted vote	0.469	0.316	0.221	0.164	0.347	0.212
Full Strategy	**0.487**	**0.332**	**0.234**	**0.175**	**0.362**	**0.225**
Loss weighting: Fixed weights	0.471	0.318	0.222	0.165	0.348	0.213
Loss weighting: GradNorm	0.480	0.327	0.230	0.172	0.357	0.221
No rule supervision ($\mathcal{L}_{\text{rule}}$ removed)	0.435	0.285	0.197	0.145	0.322	0.192
Active sampling: Random selection	0.463	0.310	0.217	0.161	0.342	0.208
Active sampling: Entropy only	0.475	0.323	0.228	0.170	0.353	0.217
Active sampling: Diversity only (k-center)	0.469	0.317	0.222	0.165	0.348	0.212
Retrieval ablation: $N_r = 0$ (no retrieval)	0.459	0.305	0.216	0.161	0.339	0.206
Retrieval size $N_r = 10$	0.485	0.330	0.232	0.174	0.360	0.223
No LLM constrained paraphrasing	0.482	0.328	0.231	0.172	0.358	0.221
Without feedback simulator	0.476	0.324	0.228	0.170	0.354	0.218

4.3 Experimental Results and Comprehensive Ablations

As shown in Table 1, NeuroSymb-MRG outperforms all baselines across language and semantic metrics on IU X-ray and MIMIC-CXR. Ablation results in Table 2 and Table 3 underscore the necessity of each component: the differentiable logic pipeline remains the primary performance driver, while the integration of learnable gating, uncertainty-aware decoding, and active-learning strategies (sampling and feedback) ensures optimal data efficiency and minimizes overconfident errors.

5 Conclusion

We presented NeuroSymb-MRG, a neuro-symbolic framework that integrates differentiable abductive reasoning with retrieval-augmented generation and active uncertainty minimization to produce structured and interpretable radiology reports. The proposed architecture produces interpretable rule-level explanations and leverages rule-level uncertainty to focus targeted training where it matters most. Empirical results show improved structural coherence and interpretability and standard automatic metrics on widely used public benchmarks. Future work will explore clinical knowledge integration, multi-institutional validation, and efficacy metrics beyond n-gram overlap.

Acknowledgement. This work was supported by the Foshan Higher Education Foundation (No. BKBS202203), the Clinical Medicine Research Pilot Project of the First People's Hospital of Foshan (No. FSYYY202503006), the Macau FDCT (Grant Nos. 0019/2025/RIB1 and 0024/2025/AIJ), the University of Macau research grants (No. EF2025-00287-FST), and the ZUMRI-DeepFuture Technology Joint Lab (Project No. CP-104-2025(2)).

Disclosure of Interest. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agishev, R., Zimmermann, K.: Fusionforce: End-to-end differentiable neural-symbolic layer for trajectory prediction. arXiv preprint arXiv:2502.10156 (2025)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
3. Chen, W., Liu, Y., Wang, C., Zhu, J., Li, G., Liu, C.L., Lin, L.: Cross-modal causal intervention for medical report generation. arXiv preprint arXiv:2303.09117 (2023)
4. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers). pp. 5904–5914 (2021)

5. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). pp. 1439–1449 (2020)
6. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10578–10587 (2020)
7. Cotnareanu, J., Chetelat, D., Zhang, Y., Coates, M.: A balanced neuro-symbolic approach for commonsense abductive logic. arXiv preprint arXiv:2601.18595 (2026)
8. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
9. Hirsch, E., Dawidowicz, G., Tal, A.: Medrat: Unpaired medical report generation via auxiliary tasks. In: European Conference on Computer Vision. pp. 18–35. Springer (2024)
10. Jiang, P., Ouyang, S., Jiao, Y., Zhong, M., Tian, R., Han, J.: Retrieval and structuring augmented generation with large language models. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. pp. 6032–6042 (2025)
11. Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 2607–2615 (2024)
12. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
13. Junhwan, C., Seokmin, O., Joongmoo, B.: Uncertainty estimation in avo inversion using bayesian dropout based deep learning. Journal of Petroleum Science and Engineering **208**, 109288 (2022)
14. Li, C., Hou, J., Shi, Y., Hu, J., Zhu, X.X., Mou, L.: Multimodal large language models for medical report generation via customized prompt tuning. arXiv preprint arXiv:2506.15477 (2025)
15. Liang, B., Wang, Y., Tong, C.: Ai reasoning in deep learning era: From symbolic ai to neural–symbolic ai. Mathematics **13**(11), 1707 (2025)
16. Liang, Y., Huang, F., Qiu, Z., Shu, X., Liu, Q., Yuan, D.: Uncertainty and diversity-based active learning for uav tracking. Neurocomputing **639**, 130265 (2025)
17. Liu, F., Ge, S., Wu, X.: Competence-based multimodal curriculum learning for medical report generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 3001–3012 (2021)
18. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13753–13762 (2021)
19. Liu, F., Yin, C., Wu, X., Ge, S., Zhang, P., Sun, X.: Contrastive attention for automatic chest x-ray report generation. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021. pp. 269–280 (2021)
20. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 375–383 (2017)

21. Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M.: Progressive transformer-based generation of radiology reports. In: Findings of the association for computational linguistics: EMNLP 2021. pp. 2824–2832 (2021)
22. Qin, H., Song, Y.: Reinforced cross-modal alignment for radiology report generation. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 448–458 (2022)
23. Rani, V., Kumar, M., Gupta, A., Sachdeva, M., Mittal, A., Kumar, K.: Self-supervised learning for medical image analysis: a comprehensive review. Evolving Systems **15**(4), 1607–1633 (2024)
24. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7008–7024 (2017)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
26. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
27. Wang, J., Bhalerao, A., He, Y.: Cross-modal prototype driven network for radiology report generation. In: European Conference on Computer Vision. pp. 563–579. Springer (2022)
28. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology **1**(3), 100033 (2023)
29. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 655–664. Springer (2022)
30. Wasilewski, P., Nguy, C.D.: Deep differentiable logic gate networks based on fuzzy zadeh's t-norm. In: Polish Conference on Artificial Intelligence. pp. 57–70. Springer (2025)
31. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 72–82. Springer (2021)
32. Zhang, K., Jiang, H., Zhang, J., Huang, Q., Fan, J., Yu, J., Han, W.: Semi-supervised medical report generation via graph-guided hybrid feature consistency. IEEE Transactions on Multimedia **26**, 904–915 (2023)